

# Great Expectations: EM Algorithms for Discretely Observed Linear Birth-Death-Immigration Processes

Charles R. Doss<sup>1</sup>, Marc A. Suchard<sup>2</sup>, Ian Holmes<sup>3</sup>, Midori Kato-Maeda<sup>4</sup>, and Vladimir N. Minin<sup>1,\*</sup>

<sup>1</sup>Department of Statistics, University of Washington, Seattle, WA 98195, USA

<sup>2</sup>Departments of Biomathematics, Biostatistics, and Human Genetics,  
University of California, Los Angeles, CA 90095, USA

<sup>3</sup>Department of Bioengineering and Biophysics Graduate Group,  
University of California, Berkeley, CA 94720, USA

<sup>4</sup>Department of Medicine, University of California, San Francisco, CA 94143, USA

## Abstract

Estimating parameters of continuous-time linear birth-death-immigration processes, observed discretely at unevenly spaced time points, is a recurring theme in statistical analyses of population dynamics. Viewing this task as a missing data problem, we develop two novel expectation-maximization (EM) algorithms. When birth rate is zero or immigration rate is either zero or proportional to the birth rate, we use Kendall's generating function method to reduce the E-step of the EM algorithm, as well as calculation of the Fisher information, to one dimensional integration. This reduction results in a simple and fast implementation of the EM algorithm. To tackle the unconstrained birth and immigration rates, we extend a direct sampler for finite-state Markov chains and use this sampling procedure to develop a Monte Carlo EM algorithm. We test our algorithms on simulated data and then use our new methods to explore the birth and death rates of a transposable element in the genome of *Mycobacterium tuberculosis*, the causative agent of tuberculosis.

## 1 Introduction

Linear birth-death-immigration (BDI) processes provide useful building blocks for modeling population dynamics in ecology (Nee, 2006), molecular evolution (Thorne et al., 1991), and epidemiology (Gibson and Renshaw, 1998) among many other areas. Although Keiding (1975) has extensively studied inference for fully observed continuous-time BDI processes, more often such processes are not observed completely, posing interesting computational problems for statisticians. Here, we use applied probability tools to develop new, efficient implementations of the expectation-maximization (EM) algorithm for fitting discretely observed BDI processes.

We assume that we observe one or multiple independent BDI trajectories at fixed, possibly irregularly spaced, time points. Holmes (2005) proposed an EM algorithm for such discretely observed BDI processes in the context of finding the most optimal alignment of multiple genomic sequences. Of course, the EM algorithm is not the only way to find maximum likelihood estimates (MLEs) of discretely observed BDI parameters, as demonstrated by Thorne et al. (1991), who initially proposed the BDI model for the sequence alignment problem. However, Holmes (2005) argues that the EM algorithm's simplicity and robustness make this method attractive for large-scale bioinformatics applications.

Computing expectations of the complete-data log-likelihood, needed for executing an EM algorithm, can be challenging, especially if the complete-data were generated by a continuous-time stochastic process. When complete data are generated by a finite state-space continuous-time Markov chain (CTMC), these expectations can be computed efficiently (Lange, 1995; Holmes and Rubin, 2002). Although the BDI process is also a CTMC, the infinite state-space of the process prohibits us from using these computationally efficient methods. Holmes (2005) considers a BDI model with the immigration rate either zero or proportional to the birth rate. Under this restriction, the complete-data likelihood belongs to the exponential family, which means that the expected complete-data log-likelihood is a linear combination of expected sufficient statistics of the complete data. Holmes (2005) computes these expectations of the sufficient statistics by numerically

solving a system of coupled non-linear ordinary differential equations (ODEs). For this restricted immigration model and for the death-immigration model we develop a new computationally efficient method for computing the expected sufficient statistics. Our method combines ideas from Kendall (1948) and Lange (1982) and reduces computations of expected sufficient statistics to one-dimensional integration, a computational task that is much simpler than solving a system of nonlinear ODEs. We develop a similar integration method to compute the observed Fisher information matrix via Louis (1982)'s formula. Moreover, for the death-immigration model and for Holmes (2005)'s sequence alignment model, we derive the expectations of the complete-data sufficient statistics in closed form.

When rates of the BDI model are not constrained, the infinite-dimensional vector of sufficient statistics precludes us from applying the above integration technique. Therefore, we resort to Monte Carlo (MC) estimation of the expected complete-data log-likelihood. Our MC-EM algorithm requires sampling BDI trajectories over a finite time-interval conditional on the observed states of the process at the end-points of the interval. Golinelli (2000) previously accomplished this task via reversible jump Markov chain Monte Carlo (MCMC). Instead of this MCMC procedure, we adopt and extend Hobolth (2008)'s direct sampler for finite state-space CTMCs to simulate end-point conditioned BDI trajectories exactly. While the resulting MC-EM algorithm is slower than our integration method for the restricted immigration BDI model, the algorithm still performs well on moderately-sized problems.

We test our two EM algorithms on simulated data. We then turn to a problem of estimating birth and death rates of the transposable element *IS6110* in *Mycobacterium tuberculosis*, the causative bacterial agent of most tuberculosis in humans. Estimating these rates is an important task in molecular epidemiology, because researchers use *IS6110* genotypes to group infected individuals into epidemiological clusters (Small et al., 1994). Rosenberg et al. (2003) use serially sampled *IS6110* genotypes from *M. tuberculosis* infected patients to estimate *IS6110* birth and death rates. These authors proposed an approximate likelihood method to accomplish this estimation. We revisit this problem using our EM algorithm and compare our results with Rosenberg et al. (2003)'s approximation. We also examine differences in birth and death rates among three main lineages of *M. tuberculosis*.

## 2 The EM Algorithm

We start with a continuous-time homogeneous linear BDI process  $\{X_t\}$  with birth rate  $\lambda \geq 0$ , death rate  $\mu \geq 0$ , and immigration rate  $\nu \geq 0$ . We assume that we observe the process,  $X_t = 0, 1, \dots$ , at  $n+1$  distinct times,  $0 = t_0 < t_1 < \dots < t_n$ . We denote our data vector by  $\mathbf{Y} = (X_{t_0}, \dots, X_{t_n})$  and the parameter vector by  $\boldsymbol{\theta} = (\lambda, \mu, \nu)$ . In addition to the full BDI model, we consider a restricted immigration model, in which we require  $\nu = \beta\lambda$  for some known constant  $\beta$ , and a death-immigration model, where we set  $\lambda = 0$ .

We are interested in computing the MLEs of the parameters,  $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} l_o(\mathbf{Y}; \boldsymbol{\theta})$ , where

$$l_o(\mathbf{Y}; \boldsymbol{\theta}) := \sum_{i=0}^{n-1} \log p_{X_{t_i}, X_{t_{i+1}}}(t_{i+1} - t_i; \boldsymbol{\theta}) \quad (1)$$

is the observed-data log-likelihood and  $p_{i,j}(t; \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(X_t = j | X_0 = i)$ ,  $i, j = 0, 1, \dots$ , are the transition probabilities of the BDI process. These transition probabilities can be calculated either using the generating function derived by Kendall (1948) or via the orthogonal polynomial representation of Karlin and McGregor (1958). Despite the explicit algebraic nature of the orthogonal polynomials, the latter method can be numerically unstable and the generating function method is often preferred (Sehl et al., 2009). Although one can maximize the likelihood  $l_o(\mathbf{Y}; \boldsymbol{\theta})$  using standard off-the-shelf optimization algorithms, such generic algorithms can be problematic when one needs to analyze multiple data sets without manual tuning of the algorithms and when the BDI rates are functions of a high dimensional parameter vector. As an alternative to generic optimization, we develop EM algorithms, known for their robustness and ability to cope with high

dimensional optimization, to maximize the observed data likelihood under BDI models. (Dempster et al., 1977).

Complete data in our case consist of the BDI trajectories  $\{X_t\}$ , observed continuously during the interval  $[t_0, t_n]$ . Let  $l_c$  be the log-likelihood of the complete data. To execute an EM algorithm we need to be able to compute  $E_{\theta'} [l_c(\{X_t\}; \theta) | \mathbf{Y}]$  (the E-step) and to maximize this expectation over  $\theta$  (the M-step). We develop separate algorithms for implementing these E- and M-steps for the restricted immigration/death-immigration and the full BDI models, because the two classes of models differ in the way the complete data collapse into sufficient statistics.

## 2.1 Restricted Immigration and Death-Immigration Models

Since the BDI process is a CTMC, the log-likelihood of complete data is

$$l_c(\{X_t\}; \theta) = - \sum_{i=0}^{\infty} d(i) [i(\lambda + \mu) + \nu] + \sum_{i=0}^{\infty} [n_{i,i+1} \log(i\lambda + \nu) + n_{i,i-1} \log(i\mu)] + \text{const}, \quad (2)$$

where  $d(i)$  is the total time spent in state  $i$  and  $n_{i,j}$  is the number of jumps from state  $i$  to state  $j$  during the interval  $[t_0, t_n]$  (Guttorp, 1995). Replacing  $\nu$  with  $\beta\lambda$  in the above equation, we arrive at the complete-data log-likelihood for the restricted immigration model,

$$l_c(\{X_t\}; \lambda, \mu) = -R_{t_n}(\lambda + \mu) - t_n\beta\lambda + N_{t_n}^+ \log \lambda + N_{t_n}^- \log \mu + \text{const}, \quad (3)$$

where the number of jumps up  $N_{t_n}^+ := \sum_{i \geq 0} n_{i,i+1}$ , the number of jumps down  $N_{t_n}^- := \sum_{i \geq 0} n_{i,i-1}$ , and the total particle-time  $R_{t_n} := \int_{t_0}^{t_n} X_s ds = \sum_{i=0}^{\infty} id(i)$  are sufficient statistics. Similarly, if we set  $\lambda = 0$  (death-immigration model), then

$$l_c(\{X_t\}; \mu, \nu) = -R_{t_n}\mu - t_n\nu + N_{t_n}^+ \log \nu + N_{t_n}^- \log \mu + \text{const}, \quad (4)$$

Equations (3) and (4) show that, for the E-step, the only expectations we need are  $U_{\theta, \mathbf{Y}} := E_{\theta} [N_{t_n}^+ | \mathbf{Y}]$ ,  $D_{\theta, \mathbf{Y}} := E_{\theta} [N_{t_n}^- | \mathbf{Y}]$ , and  $P_{\theta, \mathbf{Y}} := E_{\theta} [R_{t_n} | \mathbf{Y}]$ . Using the Markov property and additivity of expectations, we break these expectations into sums of expectations of the numbers of jumps up and down and the total particle time during each time interval  $[t_k, t_{k+1}]$ , conditional on  $X_{t_k}$  and  $X_{t_{k+1}}$ . Homogeneity of the BDI models suggests that in order to complete the E-step of the EM algorithm, we need to be able to calculate

$$\begin{aligned} U_{i,j}(t) &= E(N_t^+ | X_0 = i, X_t = j), \\ D_{i,j}(t) &= E(N_t^- | X_0 = i, X_t = j), \text{ and} \\ P_{i,j}(t) &= E(R_t | X_0 = i, X_t = j). \end{aligned} \quad (5)$$

Following Minin and Suchard (2008), we choose to work with restricted moments

$$\begin{aligned} \tilde{U}_{i,j}(t) &= E(N_t^+ 1_{\{X_t=j\}} | X_0 = i), \\ \tilde{D}_{i,j}(t) &= E(N_t^- 1_{\{X_t=j\}} | X_0 = i), \text{ and} \\ \tilde{P}_{i,j}(t) &= E(R_t 1_{\{X_t=j\}} | X_0 = i), \end{aligned} \quad (6)$$

that we can divide by transition probabilities  $p_{ij}(t)$  to recover the conditional expectations (5). In order to compute the restricted moments, we first consider the joint generating function

$$H_i(u, v, w, s, t) := E(u^{N^+} v^{N^-} e^{-wR} s^{X_t} | X_0 = i), \quad (7)$$

where  $0 \leq u, v, s \leq 1$  and  $w \geq 0$ . Partial derivatives of this function,

$$\begin{aligned} G_i^+(t, s) &= \frac{\partial H_i(u, 1, 0, s, t)}{\partial u} \Big|_{u=1} = \sum_{j=0}^{\infty} s^j \sum_{n=0}^{\infty} n \Pr(N_t^+ = n, X_t = j) = \sum_{j=0}^{\infty} \tilde{U}_{i,j}(t) s^j, \\ G_i^-(t, s) &= \frac{\partial H_i(1, v, 0, s, t)}{\partial v} \Big|_{v=1} = \sum_{j=0}^{\infty} s^j \sum_{n=0}^{\infty} n \Pr(N_t^- = n, X_t = j) = \sum_{j=0}^{\infty} \tilde{D}_{i,j}(t) s^j, \text{ and} \\ G_i^*(t, s) &= \frac{\partial H_i(1, 1, w, s, t)}{\partial w} \Big|_{w=0} = - \sum_{j=0}^{\infty} s^j \int_0^{\infty} x d\Pr(R_t \leq x, X_t = j) = - \sum_{j=0}^{\infty} \tilde{P}_{i,j}(t) s^j \end{aligned} \quad (8)$$

are power series with coefficients  $\tilde{U}_{i,j}(t)$ ,  $\tilde{D}_{i,j}(t)$ , and  $-\tilde{P}_{i,j}(t)$  respectively. Therefore, if we can compute  $G_i^+(t, s)$ ,  $G_i^-(t, s)$ , and  $G_i^*(t, s)$  for every possible  $t$  and  $s$ , then we should be able to recover coefficients of the corresponding power series via differentiation or integration. Numerical evaluation of partial derivatives (8) is straightforward if we can compute finite differences of  $H_i(u, v, w, s, t)$ . Remarkably,  $H_i(u, v, w, s, t)$  is available in closed form, as we demonstrate in the theorem below, so one can even obtain derivatives (8) analytically.

**Theorem 1.** *Let  $\{X_t\}$  be a linear BDI process with parameters  $\lambda \geq 0$ ,  $\mu \geq 0$ , and  $\nu \geq 0$ . Over the interval  $[0, t]$ , let  $N_t^+$  be the number of jumps up,  $N_t^-$  be the number of jumps down, and  $R_t$  be the total particle-time. Then  $H_i(u, v, w, s, t) = E(u^{N_t^+} v^{N_t^-} e^{-wR_t} s^{X_t} | X_0 = i)$  satisfies the following partial differential equation:*

$$\frac{\partial}{\partial t} H_i = [s^2 u \lambda - (\lambda + \mu + w)s + v\mu] \frac{\partial}{\partial s} H_i + \nu(us - 1)H_i, \quad (9)$$

subject to initial condition  $H_i(u, v, w, s, 0) = s^i$ . The Cauchy problem defined by equation (9) and the initial condition has a unique solution. When  $\lambda > 0$ , the solution is

$$H_i(u, v, w, s, t) = \left( \frac{\alpha_1 - \alpha_2 \frac{s-\alpha_1}{s-\alpha_2} e^{-\lambda(\alpha_2-\alpha_1)rt}}{1 - \frac{s-\alpha_1}{s-\alpha_2} e^{-\lambda(\alpha_2-\alpha_1)rt}} \right)^i \left( \frac{\alpha_1 - \alpha_2}{s - \alpha_2 - (s - \alpha_1) e^{-\lambda(\alpha_2-\alpha_1)rt}} \right)^{\frac{\nu}{\lambda}} e^{-\nu(1-u\alpha_1)t}, \quad (10)$$

where  $\alpha_i = \frac{\lambda + \mu + w \mp \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu\nu}}{2\lambda u}$  for  $i = 1, 2$ . When  $\lambda = 0$ , the solution is

$$H_i(u, v, w, s, t) = \left( s e^{-(\mu+w)t} - \frac{v\mu(e^{-(\mu+w)t} - 1)}{\mu + w} \right)^i e^{\frac{\nu u[v\mu - (\mu+w)s](e^{-(\mu+w)t} - 1)}{(\mu+w)^2} + \nu \left( \frac{uv\mu}{\mu+w} - 1 \right) t}.$$

*Proof.* Our proof, detailed in Appendix A, is a generalization of Kendall (1948)'s derivation of the generating function of  $X_t$ .  $\square$

Having  $H_i$  in closed form gives us access to functions  $G_i^+$ ,  $G_i^-$ , and  $G_i^*$ , so we are left with the task of recovering coefficients of these power series. One way to accomplish this task is to differentiate the power series repeatedly, e.g.  $\tilde{U}_{i,j}(t) = \frac{1}{j!} \frac{\partial^j G_i^+(s, t)}{\partial s^j} \Big|_{s=0}$ . In Appendix C, we demonstrate that for the death-immigration model and Holmes (2005)'s restricted BDI model, these derivatives can be found analytically. In general, repeated differentiation of  $G_i^+$ ,  $G_i^-$ , and  $G_i^*$  needs to be done numerically, making this method impractical. Instead, we extend  $G_i^+(t, \cdot)$ ,  $G_i^-(t, \cdot)$ , and  $G_i^*(t, \cdot)$  to the boundary of a unit circle in the complex plane by the change of variables  $s = e^{2\pi iz}$  ( $i$  in this context is the imaginary number  $\sqrt{-1}$ , not the initial state of the BDI process). For example,

$$G_i^+(t, e^{2\pi iz}) = \sum_{j=0}^{\infty} \tilde{U}_{i,j}(t) e^{2\pi i j z}$$

is a periodic function in  $z$ , which means that  $\tilde{U}_{l,j}(t)$  are Fourier coefficients of this periodic function. Therefore, we can use the Riemann approximation to the Fourier transform integral to obtain

$$\tilde{U}_{l,j}(t) = \int_0^1 G_l^+(t, e^{2\pi is}) e^{-2\pi ibs} ds \approx \frac{1}{K} \sum_{k=0}^{K-1} G_l^+(t, e^{2\pi ik/K}) e^{-2\pi ibk/K},$$

for some suitably large  $K$ . The Fast Fourier Transform (FFT) (Henrici, 1979) can be applied to compute quickly multiple Fourier coefficients (Lange, 1982; Dorman et al., 2004; Suchard et al., 2008). We do not, however, use FFT in our algorithm, because for a particular time interval length  $t$ , we almost always need to compute  $\tilde{U}_{i,j}(t)$ ,  $\tilde{D}_{i,j}(t)$ ,  $\tilde{P}_{i,j}(t)$  for only one value of  $j$ .

To complete the M-step at the  $k$ th iteration of the EM algorithm, we differentiate equation (3) with respect to  $\lambda$  and  $\mu$  to obtain

$$\hat{\mu}_{k+1} = \frac{D_{\theta_k, \mathbf{Y}}}{P_{\theta_k, \mathbf{Y}}} \text{ and } \hat{\lambda}_{k+1} = \frac{U_{\theta_k, \mathbf{Y}}}{P_{\theta_k, \mathbf{Y}} + \beta t_n}$$

for the restricted immigration model. Similarly, we use equation (4) to find

$$\hat{\mu}_{k+1} = \frac{D_{\theta_k, \mathbf{Y}}}{P_{\theta_k, \mathbf{Y}}} \text{ and } \hat{\nu}_{k+1} = \frac{U_{\theta_k, \mathbf{Y}}}{t_n}$$

for the death-immigration model.

We obtain the observed Fisher information via Louis (1982)'s formula:

$$\hat{I}_{\mathbf{Y}}(\hat{\theta}) = E_{\hat{\theta}} \left[ -\ddot{l}_c(\{X_t\}; \hat{\theta}) | \mathbf{Y} \right] - E_{\hat{\theta}} \left[ \dot{l}_c(\{X_t\}; \hat{\theta}) \dot{l}_c(\{X_t\}; \hat{\theta})' | \mathbf{Y} \right],$$

where  $\dot{l}_c$  is the gradient and  $\ddot{l}_c$  is the Hessian of the complete-data log-likelihood. This requires the calculation of the conditional cross-product means,  $E[N_t^+ N_T^- | \mathbf{Y}]$ ,  $E[N_t^+ R_T | \mathbf{Y}]$ ,  $E[N_t^- R_T | \mathbf{Y}]$ , and the conditional second moments of  $N_T^+$ ,  $N_T^-$ , and  $R_T$ . The derivation of the information in terms of these moments is in Appendix B. These conditional second- and cross-moments, as well as,  $P_{\mathbf{Y}}$  and  $D_{\mathbf{Y}}$ , can be computed in analogous fashion to  $U_{\mathbf{Y}}$  above, using the joint generating function (10).

## 2.2 Full BDI Model

The complete-data log-likelihood of the full BDI model, up to an additive constant, is

$$l_c(\{X_t\}; \theta) = -[R_{t_n}(\lambda + \mu) + t_n \nu] + \sum_{i=0}^{\infty} n_{i,i+1} \log(i\lambda + \nu) + \log(\mu) N_{t_n}^-. \quad (11)$$

Thus the sufficient statistics are  $\{n_{i,i+1}\}_{i \geq 0}$ ,  $N_{t_n}^-$ , and  $R_{t_n}$ . The technique we developed for the restricted immigration case does not apply to  $E(n_{i,i+1} | \mathbf{Y})$  for all  $i \geq 0$ , because we can not derive the joint generating function for state-specific jumps up,  $n_{i,i+1}$  for  $i \geq 0$ , in closed form. Moreover, even if could calculate these expectations, the problem of evaluating the infinite sum in (11) would remain. Therefore, in the E-step of the EM algorithm, we resort to Monte Carlo to compute the expected complete-data log-likelihood for the full BDI model. We apply the ascent-based MC-EM method of Caffo et al. (2005), which dynamically adjusts the number of Monte Carlo simulations to guarantee the ascent property of the EM algorithm. In order to approximate expectations of the sufficient statistics via Monte Carlo, we adapt Hobolth (2008)'s direct sampling method for finite state space CTMCs conditioned on end points to simulating full linear BDI process trajectories conditioned on end points.

Hobolth (2008)'s direct sampling is a recursive algorithm that generates realizations of  $\{X_t\}_{t=0}^T$ , conditional on  $X_0 = a$  and  $X_T = b$ . Let  $p_i = \int_0^T f_i(t) dt$  be the probability that there is a jump to state  $i$  before time  $T$ , where

$$f_i(t) = e^{-\lambda_a t} \lambda_{a,i} p_{i,b}(T-t) / p_{a,b}(T), \quad (12)$$



$i$  is  $a + 1$  or  $a - 1$ ,  $\lambda_{a,a+1} = \lambda a + \nu$ , and  $\lambda_{a,a-1} = \mu a$ . We start the algorithm by using the probabilities  $(1 - p_{a+1} - p_{a-1}, p_{a+1}, p_{a-1})$  to randomly decide whether not to jump, to jump up, or to jump down. If there are no jumps, then necessarily we started with  $a = b$  and the algorithm ends; if there is a jump up or down we use the inverse cumulative distribution function (CDF) method to simulate the time at which the jump occurs using CDFs  $F_i(t) = \int_0^t f_i(x)/p_i dx$ , for  $i = a + 1, a - 1$ . We compute and invert these CDFs numerically using quick to compute transition probability formulae from Karlin and McGregor (1958). After simulating the first jump time,  $\tau$ , over the interval  $[0, T]$ , we repeat this process by setting  $X_0 = i$ , and  $T$  to  $T - \tau$  until  $T - \tau < 0$  when the algorithm terminates. Because we can simulate very rapidly from the BDI process without conditioning on the ending state, accept-reject sampling also provides an effective method of simulating the conditional BDI process when the outcomes we condition on are moderately likely. When they are very rare, accept-reject sampling can be exceedingly slow, and Hobolth (2008)'s method is necessary.

In the M-step of the EM algorithm, the complete-data log-likelihood (11) can be written as  $l_c(\{X_t\}; \theta) = l_1(\{X_t\}; \mu) + l_2(\{X_t\}; \lambda, \nu)$ , making it possible to maximize separately over  $\mu$  and over  $(\lambda, \nu)$ . Maximizing over  $\mu$ , we get

$$\hat{\mu}_{k+1} = \frac{U_{\theta_k, \mathbf{Y}}}{P_{\theta_k, \mathbf{Y}}}.$$

Maximizing function (11) over  $(\lambda, \nu)$  is more difficult, as there is no obvious analytic solution. First, we convert this 2-dimensional optimization problem into a 1-dimensional one by noticing that setting the derivatives with respect to  $\lambda$  and  $\nu$  to 0 and then summing the two equations together yields

$$-P_{\theta_k, \mathbf{Y}} \hat{\lambda} - T \hat{\nu} + U_{\theta_k, \mathbf{Y}} = 0. \quad (13)$$

We can then write, for instance,  $\nu(\lambda) = \frac{U_{\theta_k, \mathbf{Y}} - P_{\theta_k, \mathbf{Y}} \lambda}{T}$ , plug this expression into the likelihood function (11) and apply the Newton-Raphson algorithm to maximize this function with respect to  $\lambda$ . Using the optimal value  $\hat{\lambda}$ , we recover  $\hat{\nu} = \nu(\hat{\lambda})$ . Because our Monte Carlo E-step takes most of the computing time, we allow Newton-Raphson to take as many iterations as it needs to converge under a prespecified tolerance; Newton-Raphson algorithm generally only takes between 2 and 5 iterations to converge in our experience.

## 3 Results

### 3.1 Simulations

To test our methods, we simulate data from a restricted immigration BDI model with  $\lambda = .07$ ,  $\mu = .12$  and  $\beta = 1.2$ . The parameters are chosen to resemble, but not exactly match, the dynamics of our biological example, discussed in the next subsection. We simulate 100 independent processes starting from initial states drawn uniformly between 1 and 15. From each process we collect at least two observations. We place observation times uniformly between 0 and 30. Table 1 gives some summary statistics for the simulated data.

First, we assume the restricted immigration model and apply our EM algorithm for this model with initial parameter values of 0.2 for both  $\lambda$  and  $\mu$ . We tested other choices of starting values, but the algorithm was not sensitive to them. Next, we pretend that we do not know that the data were generated under the restricted immigration model and fit the full BDI model to the simulated data using our MC-EM algorithm, starting each parameter,  $\lambda$ ,  $\mu$ , and  $\nu$ , at 0.2. The results of fitting these two models are shown in Table 2. As expected, estimation is more precise for the restricted immigration model. In the full BDI model,  $\nu$  is the most difficult to identify unless there are many observations starting very close to 0, so the variance of  $\hat{\nu}$  tends to be large.

Value	Simulated Data	IS6110 Data
Number of Intervals	387	252
Average Interval Length	5	0.35
Number of Individuals	100	196
Number of Intervals with an Increase	78	14
Average Increase given an Increase	1.5	1
Number of Intervals with a Decrease	190	14
Average Decrease given a Decrease	2.5	1.2
Number of Intervals with No Change	119	224
Mean Starting State	5.5	11
Standard Deviation of Starting State	3.8	5.3
Total Length of Time	1947	89

Table 1: Summary statistics for the simulated and *M. tuberculosis* IS6110 data.

	$\lambda$	$\mu$	$\nu$
Restricted Immigration	0.067 (0.052, 0.081)	0.12 (0.10, 0.14)	
Full Model	0.057 (0.039, 0.074)	0.12 (0.099, 0.14)	0.11 (0.058, 0.16)
True Value	0.07	0.12	0.084

Table 2: Results of EM algorithms applied to simulated data under the restricted immigration and full BDI models. Reported values are maximum likelihood estimates and 95% confidence intervals.

### 3.2 Comparison with the Frequent Monitoring Method

We compare our EM algorithm for computing the actual MLE to the frequent monitoring (FM) method of Rosenberg et al. (2003) for computing the MLE of an approximate likelihood. In the FM method, Rosenberg et al. (2003) assume that if the starting and ending values of the birth-death process are equal for a particular interval, then no jumps occurred in this interval. Further, if the difference between the starting and ending values is  $-1$  or  $1$ , then exactly one jump up or exactly one jump down must have occurred respectively. The authors exclude all observed intervals, for which starting and ending values differ by more than one unit. Let  $i$  be the starting state for an interval,  $t$  the length of the interval, and  $\lambda_i = i(\lambda + \mu)$ . Then the corresponding probabilities for the three possible events are  $e^{-\lambda_i u}$ ,  $\frac{i\lambda}{\lambda_i}(1 - e^{-\lambda_i u})$ , and  $\frac{i\mu}{\lambda_i}(1 - e^{-\lambda_i u})$  respectively. Rosenberg et al. (2003) use this FM method to estimate rates in what is effectively a multi-state branching process, but we will compare the two methods on our restricted immigration BDI model with immigration rate constrained to be 0. We again simulate an underlying BDI process using  $\lambda = 0.07$  and  $\mu = 0.12$ . To compare the two methods, we generate three different sets of data. In each set, we generate observed states of the BDI process at a fixed constant distance  $dt$  apart. This distance varies across the data sets, taking the values .2, .4, and .6, respectively. We repeat this procedure 200 times and compute birth and death rate estimates and corresponding 95% confidence intervals using the EM algorithm and FM approximation method. We show box plots of the resulting estimates for  $\lambda$  and  $\mu$  in Figure 1. As expected, the FM estimates behave reasonably when interval lengths are small, but the approximation becomes poor as we increase the interval length. The FM method always underestimates the parameters since the method effectively undercounts the number of unobserved jumps in the BDI process. We also compute Monte Carlo estimates of coverage probabilities of the two methods, shown above the box plots in Figure 1. Not surprisingly, coverage of the 95% confidence intervals computed under the proper BDI model likelihood are very close to the promised value of 0.95. In contrast, the FM approximation-based 95% confidence intervals contain the true parameter value less than 95% for all three simulation scenarios.

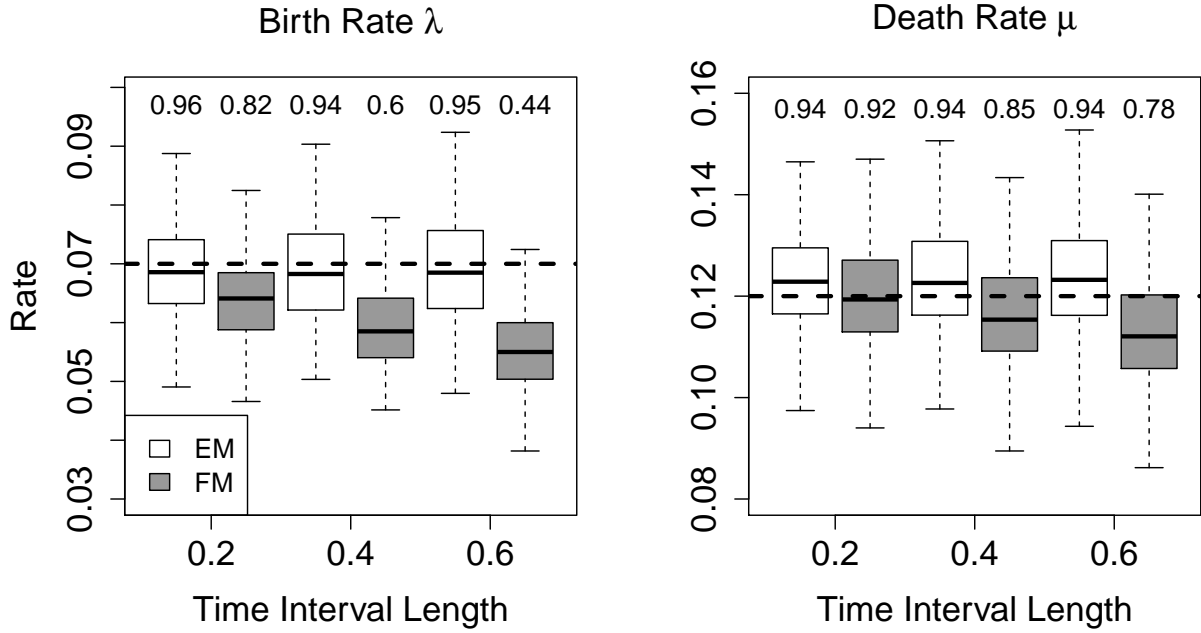


Figure 1: Box plots of birth (left panel) and death (right panel) rate estimates, obtained from 200 simulated data sets using the EM algorithm and frequent monitoring (FM) method. The true parameter values, used in data simulations, are marked by the horizontal dashed lines. Above the box plots, we show Monte Carlo estimates of coverage probabilities that the 95% confidence intervals attain.

### 3.3 *Mycobacterium tuberculosis* IS6110 Transposon

We apply our restricted immigration EM algorithm to estimation of birth and death rates of the transposon IS6110 in *M. tuberculosis* (McEvoy et al., 2007). A transposon or transposable element is a genetic sequence that can duplicate, remove itself, and jump to a new location in the genome. IS6110 is a transposon that plays an important role in epidemiological studies of tuberculosis. More specifically, the number and locations of IS6110 elements in the *M. tuberculosis* form a genetic signature or genotype of the mycobacterium, allowing epidemiologists to draw inference about disease transmission when the same genotype is observed among patients with active tuberculosis (van Embden et al., 1993). Such genotypic comparison can translate into meaningful epidemiological inference only if the dynamics of IS6110 evolution are well understood. Therefore, accurate estimation of rates of changes of IS6110-based genotypes is critical for using these genotypes in epidemiological studies (Tanaka and Rosenberg, 2001).

We analyze data from an ongoing population-based study that includes all tuberculosis cases reported to the San Francisco Department of Public Health (Cattamanchi et al., 2006). Our data include patients with more than one *M. tuberculosis* isolate from specimens sampled more than 10 days apart and genotyped with IS6110 restriction fragment length polymorphism. We ignore genomic locations of IS6110 and assume that the transposon counts are discretely observed realizations of a BDI process, with no immigration; in particular, we assume that the patient is not reinfected with a different strain of the bacteria in the period between observations. Table 1 gives summary statistics for the data.

We plot birth and death rate estimates and their 95% confidence intervals, obtained with our EM algorithm, in Figure 2 (vertical bars labeled “All”). The starting values for the EM do not affect the results. In the analysis presented, the EM algorithm was started with parameter guesses of .05 and .05 for  $\lambda$  and  $\mu$ , respectively, and their MLEs were 0.027 and 0.031, respectively. Our estimate for  $\lambda$  is consistent with the estimate  $0.0188 \pm 0.0103$  from Rosenberg et al. (2003). Although the authors’ credible interval for  $\mu$  overlaps with



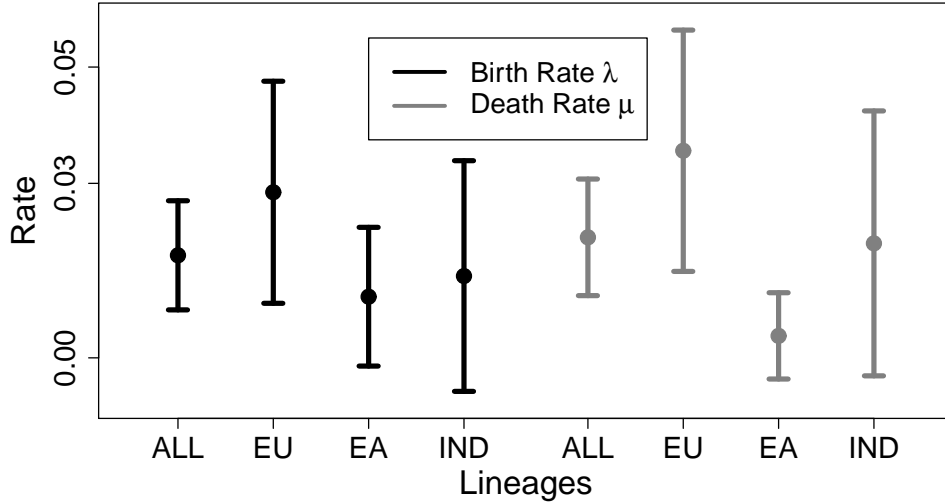


Figure 2: Point estimates and 95% confidence intervals for birth and death rate of the *IS6110* transposable element obtained from all individuals (ALL) and obtained by separately analyzing three *M. tuberculosis* lineages: European-American (EU), Indo-Oceanic (IND), and East Asian (EA).

ours, our estimate for  $\mu$  is noticeably higher than Rosenberg et al. (2003)’s estimate of  $0.0147 \pm 0.00906$ . Note from Table 1 that among the intervals with a decrease, the average count drop was by more than 1; there were 3 intervals where *IS6110* counts dropped by 2, whereas there were no interval that experienced an increase by more than 1. Thus we would expect our estimate for  $\mu$  to increase over Rosenberg et al. (2003)’s approximation, whereas that of  $\lambda$  should be similar between the two methods. We also point out that we analyze an updated version of the data analyzed by Rosenberg et al. (2003). Moreover, Rosenberg et al. (2003) use a slightly more complicated model for *IS6110* evolution, which takes into account shifts in transposon location. Given these differences in the data and the methods, consistency of our and Rosenberg et al. (2003)’s estimates is comforting.

### 3.3.1 *Mycobacterium tuberculosis* Lineage Comparison

In addition to estimation of the global birth and death rates, we separately estimate these parameters in each of the three lineages of *M. tuberculosis* observed in San Francisco. Based on genomic sequence similarity, *M. tuberculosis* is divided into six main lineages: Euro-American, East-Asian, Indo-Oceanic, East-African-Indian, West-African I and West-African II (Gagneux et al., 2006). In our lineage-specific analysis, we consider 109 individuals infected with Euro-American lineage strains, 54 individuals infected with East-Asian lineage strains, and 25 individuals infected with Indo-Oceanic lineage strains. The *M. tuberculosis* lineage-specific estimates and confidence intervals are plotted in Figure 2. Most notably, there appears to be a substantial difference between death rates of the Euro-American and East-Asian lineages. Since this is a novel result that has implications for monitoring tuberculosis with molecular genotyping, we examine the difference in death rates between lineages more closely.

The number of *IS6110* elements is a potential confounder in our analysis, because patients infected with Euro-American and East-Asian differ drastically in the number of *IS6110* elements at the beginning of the observation period. The isolates from the Euro-American lineage have between 2 and 17 *IS6110* elements, with 41 out of 109 patients having the first recorded *IS6110* count less than 6, while *IS6110* counts vary between 6 and 22 for the East-Asian isolates. Warren et al. (2002) suggest that *IS6110* genotypes with fewer

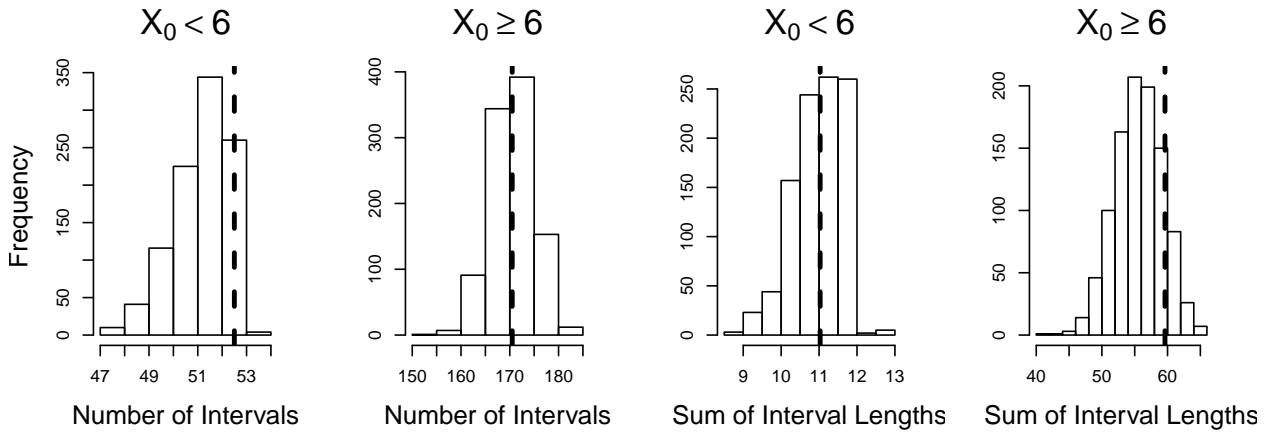


Figure 3: Low vs high count genotype analysis. Histograms of simulated numbers of intervals and sums of interval lengths are plotted for intervals with starting values less than six and greater or equal to six. The vertical dashed lines indicate the observed values of the four statistics.

than six elements have very low rate of change, because in their data, cases with no observed changes in the genotype are dominated by such low-count genotypes. However, according to our linear birth-death model, Warren et al. (2002)’s observation of low-count genotypes evolving slower than high-count genotypes is nothing but expected. To demonstrate this, we simulate 1000 datasets using our global birth and death rates and observed initial IS6110 counts for each patient. We record the number of intervals with equal starting and ending values less than six,  $n_{0,<6}$ , and equal starting and ending values greater or equal to six,  $n_{0,\geq 6}$ . We also recorded the length sum of both kinds of intervals:  $t_{0,<6}$  and  $t_{0,\geq 6}$ . In our data,  $n_{0,<6}^{\text{obs}} = 53$  and  $n_{0,\geq 6}^{\text{obs}} = 171$  with  $n_{0,<6}^{\text{obs}}/t_{0,<6}^{\text{obs}} = 4.6 > 2.8 = n_{0,\geq 6}^{\text{obs}}/t_{0,\geq 6}^{\text{obs}}$ , in agreement with Warren et al. (2002)’s analysis. Histograms of simulated values of the four statistics,  $n_{0,<6}$ ,  $n_{0,\geq 6}$ ,  $t_{0,<6}$ , and  $t_{0,\geq 6}$ , shown in Figure 3, demonstrate that our birth-death model replicates well the observed dynamics of low-count and high-count IS6110 genotypes. We conclude that our data do not provide evidence that evolutionary dynamics of low-count genotypes differ from high-count genotype dynamics. Therefore, it is unlikely that our estimated discrepancy between death rates of Euro-American and East-Asian *M. tuberculosis* lineages is caused by high percentage of low-count genotypes in the Euro-American lineage isolates.

## 4 Discussion

We propose two implementations of an EM algorithm for maximum likelihood estimation of discretely observed BDI processes. When the birth rate is zero or the immigration rate is zero or proportional to the birth rate, we show that the E-step of the EM algorithm can be reduced to computing a small number of one-dimensional Fourier transform integrals. This makes our method more efficient than Holmes (2005)’s strategy, which involves finding numerical solutions of non-linear ODEs. Moreover, in Appendix C, we demonstrate that for the Holmes (2005)’s EM algorithm and for the death-immigration model, our generating function method yields analytic formulae for the expected complete-data log-likelihood. Therefore, for these classes of models, our EM algorithm is exact, meaning that neither E-step nor M-step of the algorithm requires numerical approximations.

Our second EM algorithm uses exact sampling of end-point conditioned BDI trajectories. This sampling algorithm is a direct extension of Hobolth (2008)’s algorithm. The key difference is that Hobolth (2008) worked with finite state-space CTMCs with diagonalizable infinitesimal generators. Assuming that the spectral decomposition of the generator is available, the author was able to obtain analytic expressions for

the CDFs of waiting times until the next jump in the algorithm. Although Karlin and McGregor (1958)’s orthogonal polynomial decomposition of BDI transition probabilities is an analog of the matrix spectral decomposition, we were not able to use the orthogonal polynomials to avoid numeric integration in our exact sampling algorithm. Since this numeric integration is a major bottleneck in the algorithm, any progress in using Karlin and McGregor (1958)’s orthogonal polynomials to speed up the integration step will be worth the effort.

To simplify our presentation, we concentrated on a simple parameterization of BDI processes in terms of birth, death, and immigration rates. In many applications, rates of continuous-time Markov processes are modeled as functions of covariates  $\mathbf{z}$  (Kalbfleisch and Lawless, 1985). For example, in the BDI model we can set  $\log \lambda(\boldsymbol{\theta}_\lambda) = \mathbf{z}^t \boldsymbol{\theta}_\lambda$ ,  $\log \mu(\boldsymbol{\theta}_\mu) = \mathbf{z}^t \boldsymbol{\theta}_\mu$ ,  $\log \nu(\boldsymbol{\theta}_\nu) = \mathbf{z}^t \boldsymbol{\theta}_\nu$ , where  $\boldsymbol{\theta}_\lambda, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\nu$  are parameters of interest. Under such log-linear parameterization, the E-steps of our EM algorithms remain intact, making our numerical integration methods relevant for potentially high-dimensional statistical problems involving BDI processes. One can invoke standard modifications of the EM algorithm, such as gradient EM (Lange, 1995) and Expectation/Conditional Maximization (Meng and Rubin, 1993), to circumvent the analytical intractability of the M-step under such complex parameterizations of the BDI model.

If one works in a Bayesian framework, it is clear that the exact sampling algorithm can also be used to draw BDI trajectories from their full conditional distribution. This observation suggests a Bayesian data augmentation (BDA) MCMC algorithm, where BDI trajectories play the role of missing data. Since independent gamma priors on birth and death rates are conjugate to the complete-data likelihood of the restricted immigration BDI model, the BDA MCMC can be accomplished via a two stage Gibbs sampler (Tanner and Wong, 1987). Golinelli (2000) develops a similar BDA MCMC algorithm, except the author uses reversible jump MCMC to sample BDI trajectories. Notoriously slow mixing of reversible jump MCMC suggests that the two stage Gibbs sampler should be more efficient than Golinelli (2000)’s method. Unfortunately, in the unrestricted immigration BDI model, the prior conjugacy is lost, diminishing the appeal and simplicity of the BDA MCMC.

Finally, we would like to point out that the generating functions derived in Theorem 1 are useful not only for computing expected complete-data log-likelihoods of BDI models. For example, we are not aware of analytic formulae for expectations of sufficient statistics that do not involve the ending state of the process at time  $t$ :  $E(N_t^+ | X_0 = i)$ ,  $E(N_t^- | X_0 = i)$ , and  $E(R_t^+ | X_0 = i)$ . These expectations, useful for prediction purposes, can be obtained analytically from the generating functions in Theorem 1 (e.g.  $E(N_t^+ | X_0 = i) = \partial H_i(u, 1, 0, 1, t) / \partial u|_{u=1}$ ).

## 5 Implementation

We implemented the EM algorithm and BDA MCMC for restricted and full BDI models in the R package DOBAD (Discretely Observed Birth And Death processes), available from CRAN (CRAN, 2010) (<http://cran.r-project.org/web/packages/DOBAD/>).

## Acknowledgments

We thank Peter Guttorp for stimulating discussions and for pointing us to the work of Golinelli (2000). VNM was partially supported by the UW Royalty Research Fund and by the NSF grant No. DMS-0856099. MAS was supported by the NIH grant No. R01 GM086887. IH was supported by the NIH grant No. GM076705.

## References

- Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo expectation-maximization. Journal of the Royal Statistical Society, Series B **67**, 235–251.
- Cattamanchi, A., Hopewell, P., Gonzalez, L., Osmond, D., Masae, Kawamura, L., Daley, C., and Jasmer, R. (2006). A 13-year molecular epidemiological analysis of tuberculosis in San Francisco. The International Journal of Tuberculosis and Lung Disease **10**, 297–304(8).
- CRAN (2010). The comprehensive R archive network. <http://cran.r-project.org>.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**, 1–38.
- Dorman, K., Sincheimer, J., and Lange, K. (2004). In the garden of branching processes. SIAM Review **46**, 202–229.
- Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M., Hilty, M., Hopewell, P., and Small, P. (2006). Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences, USA **103**, 2869–2873.
- Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. IMA Journal of Mathematics Applied in Medicine & Biology **15**, 19–40.
- Golinelli, D. (2000). Bayesian Inference in Hidden Stochastic Population Processes. PhD thesis, University of Washington.
- Guttorp, P. (1995). Stochastic Modeling of Scientific Data. Chapman & Hall, London.
- Henrici, P. (1979). Fast Fourier transform methods in computational complex analysis. SIAM Review **21**, 481–527.
- Hobolth, A. (2008). A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. Journal of Computational and Graphical Statistics **17**, 138–162.
- Holmes, I. (2005). Using evolutionary expectation maximization to estimate indel rates. Bioinformatics **21**, 2294–2300.
- Holmes, I. and Rubin, G. (2002). An expectation maximization algorithm for training hidden substitution models. Journal of Molecular Biology **317**, 753 – 764.
- Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. Journal of the American Statistical Association **80**, 863–871.
- Karlin, S. and McGregor, J. (1958). Linear growth birth and death processes. Journal of Mathematics and Mechanics **7**, 643–662.
- Keiding, N. (1975). Maximum likelihood estimation in the birth-and-death process. The Annals of Statistics **3**, 363–372.
- Kendall, D. G. (1948). On the generalized “birth-and-death” process. Annals of Mathematical Statistics **19**, 1–15.

- Lange, K. (1982). Calculation of the equilibrium distribution for a deleterious gene by the finite Fourier transform. Biometrics **38**, 79–86.
- Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society, Series B **57**, 425–437.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B **44**, 226–233.
- McEvoy, C., Falmer, A., van Pittius, N., Victor, T., van Helden, P., and Warren, R. (2007). The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. Tuberculosis **87**, 393–404.
- Meng, X. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**, 267–278.
- Minin, V. N. and Suchard, M. A. (2008). Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology **56**, 391–412.
- Nee, S. (2006). Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics **37**, 1–17.
- Neuts, M. F. (1995). Algorithmic Probability: A Collection of Problems. Stochastic Modeling Series. Chapman & Hall, London.
- Rosenberg, N. A., Tsolaki, A. G., and Tanaka, M. M. (2003). Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in mycobacterium tuberculosis. Theoretical Population Biology **63**, 347 – 363.
- Sehl, M., Zhou, H., Sinsheimer, J., and Lange, K. (2009). Extinction models for cancer stem cell therapy. Technical report, UCLA, Department of Statistics.
- Small, P., Hopewell, P., Singh, S., Paz, A., Parsonnet, J., Ruston, D., Schechter, G., Daley, C., and Schoolnik, G. (1994). The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. New England Journal of Medicine **330**, 1703–1709.
- Suchard, M., Lange, K., and Sinsheimer, J. (2008). Efficiency of protein production from mRNA. Journal of Statistical Theory and Practice **2**, 173–182.
- Tanaka, M. and Rosenberg, N. (2001). Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Statistics in Medicine **20**, 2409–2420.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association **82**, 528–540.
- Thorne, J., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. Journal of Molecular Evolution **33**, 114–124.
- van Embden, J., Cave, M., Crawford, J., Dale, J., Eisenach, K., Gicquel, B., Hermans, P., Martin, C., McAdam, R., and Shinnick, T. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. Journal of Clinical Microbiology **31**, 406–409.
- Warren, R., van der Spuy, G., Richardson, M., Beyers, N., Booyesen, C., Behr, M., and van Helden, P. (2002). Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology **40**, 1277–1282.

# Appendices

## A Proof of Theorem 1

Here, we prove our main result.

*Proof.* We consider a joint measure  $V_{i,j}(n_1, n_2, x, t) = P(X_t = j, N_t^+ = n_1, N_t^- = n_2, R_t \leq x | X_0 = i)$ . For ease of notation, we will let  $\lambda_{ij}$  be the instantaneous rate of transitioning from state  $i$  to state  $j$  for the BDI process and  $\lambda_i = \sum_{j \neq i} \lambda_{ij}$ . Also, we will let  $a_i = i$  be the reward rate for  $R_t$ ; that is, for staying in state  $i$  for time  $h$ , the process  $R_t$  increases by  $ih$ . Following Neuts (1995), we start with

$$\begin{aligned} V_{ij}(n_1, n_2, x, t) &= 1_{\{i=j\}} 1_{\{x \geq a_i t\}} 1_{\{n_1=n_2=0\}} e^{-\lambda_i t} \\ &+ 1_{\{j \geq 1\}} 1_{\{n_1 \geq 1\}} \int_0^t V_{i,j-1}[n_1 - 1, n_2, x - (t-u)a_j, u] e^{-\lambda_j(t-u)} \lambda_{j-1,j} du \\ &+ 1_{\{n_2 \geq 1\}} \int_0^t V_{i,j+1}[n_1, n_2 - 1, x - (t-u)a_j, u] e^{-\lambda_j(t-u)} \lambda_{j+1,j} du, \end{aligned}$$

where  $1_{\{\cdot\}}$  is the indicator function. Next, we derive differential equations for the Laplace-Stieltjes transform  $V_{i,j}^*(n_1, n_2, w, t) = \int_0^\infty e^{-wx} dV_{i,j}(n_1, n_2, x, t)$ :

$$\begin{aligned} \frac{\partial}{\partial t} V_{ij}^*(n_1, n_2, w, t) &= -jwV_{ij}^*(n_1, n_2, w, t) - [j(\lambda + \mu) + \nu] V_{ij}^*(n_1, n_2, w, t) \\ &+ 1_{\{n_1 \geq 1\}} 1_{\{j \geq 1\}} [\lambda(j-1) + \nu] V_{i,j-1}^*(n_1 - 1, n_2, w, t) \\ &+ 1_{\{n_2 \geq 1\}} \mu(j+1) V_{i,j+1}^*(n_1, n_2 - 1, w, t). \end{aligned}$$

We now write  $H_i(u, v, w, s, t) = \sum_j h_{i,j}(u, v, w, t) s^j$  where  $h_{i,j}(u, v, w, t) := \sum_{n_1, n_2} V_{i,j}^*(n_1, n_2, w, t) u^{n_1} v^{n_2}$ . The functions  $h_{i,j}$  then satisfy

$$\begin{aligned} \frac{\partial}{\partial t} h_{ij}(u, v, w, t) &= -[j(\lambda + \mu + w) + \nu] h_{ij}(u, v, w, t) + [\lambda(j-1) + \nu] u h_{i,j-1}(u, v, w, t) 1_{\{j \geq 1\}} \\ &+ (j+1) \mu v h_{i,j+1}(u, v, w, t). \end{aligned}$$

Using this fact, we arrive at

$$\begin{aligned} \frac{\partial}{\partial t} H_i &= -s \sum_{j \geq 1} s^{j-1} j(\lambda + \mu + w) h_{ij} + \sum_{j \geq 1} s^j (-\nu) h_{i,j} + -\nu h_{i0} + s \sum_{j \geq 1} s^{j-1} u \gamma h_{i,j-1} \\ &+ \sum_{j \geq 1} s^j v (j+1) \mu h_{i,j+1} + v \mu h_{i,1} + s^2 \sum_{j \geq 1} s^{j-2} u (j-1) \lambda h_{i,j-1} \\ &= -(\lambda + \mu + w) s \frac{\partial}{\partial s} H_i - \nu H_i + s u \nu H_i + v \mu \frac{\partial}{\partial s} H_i + s^2 u \lambda \frac{\partial}{\partial s} H_i, \end{aligned}$$

which proves that  $H_i$  satisfies equation (9).

Using the method of characteristics, we solve the above PDE with initial condition  $H_i(u, v, w, s, 0) = s^i$ . When  $\lambda > 0$ , the solution is

$$H_i(u, v, w, s, t) = \left( \frac{\alpha_1 - \alpha_2 \frac{s - \alpha_1}{s - \alpha_2} e^{-\lambda(\alpha_2 - \alpha_1)rt}}{1 - \frac{s - \alpha_1}{s - \alpha_2} e^{-\lambda(\alpha_2 - \alpha_1)rt}} \right)^i \left( \frac{\alpha_1 - \alpha_2}{s - \alpha_2 - (s - \alpha_1) e^{-\lambda(\alpha_2 - \alpha_1)rt}} \right)^{\frac{\nu}{\lambda}} e^{-\nu(1-u\alpha_1)t},$$



where  $\alpha_i = \frac{\lambda + \mu + w \mp \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu w}}{2\lambda u}$ , for  $i = 1, 2$ . In the case of  $\lambda = 0$  (death-immigration model), the solution is

$$H_i(u, v, w, s, t) = \left( s e^{-(\mu+w)t} - \frac{v\mu (e^{-(\mu+w)t} - 1)}{\mu + w} \right)^i e^{\frac{\nu u [v\mu - (\mu+w)s] (e^{-(\mu+w)t} - 1)}{(\mu+w)^2} + \nu \left( \frac{uv\mu}{\mu+w} - 1 \right) t}.$$

□

## B Calculating the Observed Information

In this section, we provide details for calculating the observed information matrix. Louis (1982) shows that, in problems with incomplete observations, the observed information can be calculated as

$$\hat{I}_Y(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}(\ddot{l}(\boldsymbol{\theta}, X)|Y) - E_{\boldsymbol{\theta}}(\dot{l}(\boldsymbol{\theta}, X)\dot{l}(\boldsymbol{\theta}, X)'|Y)$$

where  $l$  is the complete-data likelihood for, in our case, either the restricted immigration or the full BDI model,  $\dot{l}$  is its gradient, and  $\ddot{l}$  is its Hessian matrix. Under the restricted immigration model with  $\nu = \beta\lambda$  we have

$$\dot{l}((\lambda, \mu), X) = \left( -R_T - \beta T + \frac{N_T^+}{\lambda}, -R_T + \frac{N_T^-}{\mu} \right)',$$

so that

$$\dot{l}((\lambda, \mu), X)\dot{l}((\lambda, \mu), X)' = \begin{pmatrix} A & B \\ B & C \end{pmatrix},$$

where  $A = R_T^2 + 2R_T T\beta - \frac{2R_T N_T^+}{\lambda} - 2\frac{T\beta N_T^+}{\lambda} + \frac{N_T^{+2}}{\lambda^2} + T^2\beta^2$ ,  $B = R_T^2 + T R_T \beta - \frac{N_T^+ R_T}{\lambda} - \frac{N_T^- R_T}{\mu} - \frac{T\beta N_T^-}{\mu} + \frac{N_T^+ N_T^-}{\lambda\mu}$ , and  $C = R_T^2 - \frac{2R_T N_T^-}{\mu} + (\frac{N_T^-}{\mu})^2$ . The Hessian is

$$\ddot{l}((\lambda, \mu), X) = - \begin{pmatrix} \frac{N_T^+}{\lambda^2} & 0 \\ 0 & \frac{N_T^-}{\mu^2} \end{pmatrix}.$$

The generating function presented in Theorem 1 can be used to compute the conditional means of all the needed cross-products and square terms in the gradient and Hessian, similarly to the first moment calculations outlined in the main text.

In the full BDI model, the score function is

$$\dot{l}((\lambda, \mu, \nu), X) = \left( -R_T + \sum_{i=0}^{\infty} \frac{i n_{i,i+1}}{i\lambda + \nu}, -R_T + \frac{N_T^-}{\mu}, -T + \sum_{i=0}^{\infty} \frac{n_{i,i+1}}{i\lambda + \nu} \right)'$$

and the Hessian is

$$\ddot{l}((\lambda, \mu, \nu), X) = - \begin{pmatrix} \sum_{i=0}^{\infty} \frac{i^2 n_{i,i+1}}{(i\lambda + \nu)^2} & 0 & \sum_{i=0}^{\infty} \frac{i n_{i,i+1}}{(i\lambda + \nu)^2} \\ 0 & \frac{N_T^-}{\mu^2} & 0 \\ \sum_{i=0}^{\infty} \frac{i n_{i,i+1}}{(i\lambda + \nu)^2} & 0 & \sum_{i=0}^{\infty} \frac{n_{i,i+1}}{(i\lambda + \nu)^2} \end{pmatrix}.$$

Since our generating function method precludes us from computing conditional expectations of some entries in the above expressions, we estimate expectations of the score and Hessian via Monte Carlo with the help of our direct sampling algorithm, described in the main text.

## C Special Cases

In this section, we show that for two important special cases of the BDI model the E-step of the EM algorithm does not require any numeric approximations.

### Death-Immigration Model

We have shown that the generating function,  $H_i(u, v, w, s, t) = E(u^{N_t^+} v^{N_t^-} e^{-wR_t} s^{X_t} | X_0 = i)$ , for the death-immigration model is

$$H_i(u, v, w, s, t) = \left( s e^{-(\mu+w)t} - \frac{v\mu (e^{-(\mu+w)t} - 1)}{\mu + w} \right)^i e^{\frac{\nu u [v\mu - (\mu+w)s] (e^{-(\mu+w)t} - 1)}{(\mu+w)^2} + \nu \left( \frac{uv\mu}{\mu+w} - 1 \right) t}.$$

Suppose we are interested in computing  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = i)$ . First, we fix  $v = 1$  and  $w = 0$ . Next, we differentiate the generating function once with respect to  $u$  and  $j$  times with respect to  $s$ , plugging in 1 and 0 respectively:

$$E(N_t^+ 1_{\{X_t=j\}} | X_0 = i) = \frac{\partial}{\partial u} \frac{\partial^j}{\partial s^j} H_i(u, 1, 0, s, t) \Big|_{u=1, s=0}$$

and

$$H_i(u, 1, 0, s, t) = [1 + e^{-\mu t}(s - 1)]^i e^{-\frac{\nu u(s-1)(e^{-\mu t}-1)}{\mu} + \nu(u-1)t} = (A + Bs)^i e^{C(u)s + D(u)},$$

where

$$\begin{aligned} A &= 1 - e^{-\mu t}, \\ B &= e^{-\mu t}, \\ C(u) &= -\frac{\nu u (e^{\mu t} - 1)}{\mu}, \\ D(u) &= \frac{\nu u (e^{\mu t} - 1)}{\mu} + \nu(u - 1)t. \end{aligned}$$

Therefore,

$$\frac{\partial}{\partial u} H_i(u, 1, 0, s, t) \Big|_{u=1} = (A + Bs)^i e^{C(1)s + D(1)} [C'(1)s + D'(1)],$$

where

$$\begin{aligned} C'(1) &= -\frac{\nu (e^{\mu t} - 1)}{\mu}, \\ D'(1) &= \frac{\nu (e^{\mu t} - 1)}{\mu} + \nu t. \end{aligned}$$

Now, the derivatives with respect to  $s$  can be recovered by expanding  $\left. \frac{\partial}{\partial u} H_i(u, 1, 0, s, t) \right|_{u=1}$  into a power series:

$$\begin{aligned}
\left. \frac{\partial}{\partial u} H_i(u, 1, 0, s, t) \right|_{u=1} &= (A + Bs)^i e^{C(1)s} e^{D(1)} [C'(1)s + D'(1)] = e^{D(1)} [C'(1)s + D'(1)] \\
&\times \left[ \sum_{m=0}^i \binom{i}{m} A^{i-m} B^m s^m \right] \left[ \sum_{k=0}^{\infty} \frac{C^k}{k!} s^k \right] \\
&= e^{D(1)} \left\{ \sum_{m=0}^{i+1} \left[ C'(1) \binom{i}{m-1} A^{i-m+1} B^{m-1} 1_{\{m \geq 1\}} + D'(1) \binom{i}{m} A^{i-m} B^m 1_{\{m \leq i\}} \right] s^m \right\} \left[ \sum_{k=0}^{\infty} \frac{C^k}{k!} s^k \right] \\
&= \sum_{n=0}^{\infty} e^{D(1)} \left\{ \sum_{k=\max\{0, n-i-1\}}^n \frac{C^k}{k!} \left[ C'(1) \binom{i}{n-k-1} A^{i-n+k+1} B^{n-k-1} 1_{\{n-k \geq 1\}} \right. \right. \\
&\quad \left. \left. + D'(1) \binom{i}{n-k} A^{i-n+k} B^{n-k} 1_{\{n-k \leq i\}} \right] \right\} s^n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
E(N_t^+ 1_{\{X_t=j\}} | X_0 = i) &= e^{D(1)} \sum_{k=\max\{0, j-i-1\}}^j \frac{C^k}{k!} \left[ C'(1) \binom{i}{j-k-1} A^{i-j+k+1} B^{j-k-1} 1_{\{j-k \geq 1\}} \right. \\
&\quad \left. + D'(1) \binom{i}{j-k} A^{i-j+k} B^{j-k} 1_{\{j-k \leq i\}} \right].
\end{aligned}$$

One can derive expectations of  $N_t^-$  and  $R_t$  in a similar fashion.

## Sequence Alignment BDI Model

Here we demonstrate that our generating function approach results in analytic formulae for the E-step in the evolutionary EM algorithm, developed by Holmes (2005). This is in contrast to the original Holmes (2005)'s implementation, which requires numerically solving a system of nonlinear ordinary differential equations. Holmes (2005)'s algorithm is based on a TKF91 model of sequence alignment evolution (Thorne et al., 1991). Instead of diving into the intricacies of this model, we refer the reader to Ian Holmes' web page (<http://biowiki.org/TkfindelModelPathSummaries>), where he poses an open problem of deriving the E-step of Holmes (2005)'s algorithm in closed form and explicitly formulates this problem in terms of the BDI process. To derive the E-step of Holmes (2005)'s algorithm in closed form, using our BDI notation, one needs to find analytic expressions of the following expectations:

1.  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 1)$ ,  $E(N_t^- 1_{\{X_t=j\}} | X_0 = 1)$ , and  $E(R_t 1_{\{X_t=j\}} | X_0 = 1)$  when  $\nu = 0$ ,
2.  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 0)$ ,  $E(N_t^- 1_{\{X_t=j\}} | X_0 = 0)$ , and  $E(R_t 1_{\{X_t=j\}} | X_0 = 0)$  when  $\nu = \lambda$ ,

We derive the analytic formulae for  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 0)$  ( $\nu = \lambda$ ) and  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 1)$  ( $\nu = 0$ ). The other expectations can be derived analogously.

1. **Objective:**  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 0)$  ( $\nu = \lambda$ ):

First,

$$E(N_t^+ 1_{\{X_t=j\}} | X_0 = 0) = \left. \frac{\partial}{\partial r} \frac{\partial^j}{\partial s^j} \mathbf{H}_0^+(r, s, t) \right|_{s=0, r=1},$$

where

$$\mathbf{H}_0^+(r, s, t) = \frac{(\alpha_1 - \alpha_2) e^{-\lambda(1-r\alpha_1)t}}{s - \alpha_2 - (s - \alpha_1) e^{-\lambda(\alpha_2 - \alpha_1)rt}} \text{ and } \alpha_{1,2} = \frac{\lambda + \mu \mp \sqrt{(\lambda + \mu)^2 - 4\lambda\mu r}}{2\lambda r}.$$

We find the formula for this partial derivative by explicit differentiation:

$$\frac{\partial^j}{\partial s^j} \mathbf{H}_0^+(r, s, t) = \frac{(-1)^j j! (\alpha_1 - \alpha_2) e^{-\lambda(1-r\alpha_1)t} (1 - e^{-\lambda(\alpha_2 - \alpha_1)rt})^j}{(s - \alpha_2 - (s - \alpha_1) e^{-\lambda(\alpha_2 - \alpha_1)rt})^{j+1}},$$

$$\left. \frac{\partial^j}{\partial s^j} \mathbf{H}_0^+(r, s, t) \right|_{s=0} = \frac{(-1)^j j! (\alpha_1 - \alpha_2) e^{-\lambda(1-r\alpha_1)t} (1 - e^{-\lambda(\alpha_2 - \alpha_1)rt})^j}{(\alpha_1 e^{-\lambda(\alpha_2 - \alpha_1)rt} - \alpha_2)^{j+1}} = \frac{A(r)}{B(r)},$$

$$\left. \frac{\partial}{\partial r} \frac{\partial^j}{\partial s^j} \mathbf{H}_0^+(r, s, t) \right|_{s=0, r=1} = \frac{A'(1)B(1) - A(1)B'(1)}{B^2(1)},$$

where

$$\begin{aligned} A(1) &= \left(1 - e^{(\lambda - \mu)t}\right)^j \left(1 - \frac{\mu}{\lambda}\right), \\ B(1) &= \left(e^{(\lambda - \mu)t} - \frac{\mu}{\lambda}\right)^{j+1}, \\ A'(1) &= \left(1 - e^{(\lambda - \mu)t}\right)^{j-1} \left[ j 2\mu t e^{(\lambda - \mu)t} + \left(1 - e^{(\lambda - \mu)t}\right) \left(\frac{\lambda^2 + \mu^2}{\lambda(\mu - \lambda)} - \mu t\right) \right], \\ B'(1) &= (j+1) \left(e^{(\lambda - \mu)t} - \frac{\mu}{\lambda}\right)^j \left(\frac{\lambda(1 + 2\mu t)}{\mu - \lambda} e^{(\lambda - \mu)t} + \frac{\mu^2}{\lambda(\mu - \lambda)}\right). \end{aligned}$$

2. **Objective:**  $E(N_t^+ 1_{\{X_t=j\}} | X_0 = 1)$  ( $\nu = 0$ ):

As before,

$$E(N_t^+ 1_{\{X_t=j\}} | X_0 = 1) = \left. \frac{\partial}{\partial r} \frac{\partial^j}{\partial s^j} \mathbf{H}_1^+(r, s, t) \right|_{s=0, r=1},$$

where

$$\mathbf{H}_1^+(r, s, t) = \frac{\alpha_1(s - \alpha_2) - \alpha_2(s - \alpha_1) e^{-\lambda(\alpha_2 - \alpha_1)rt}}{s - \alpha_2 - (s - \alpha_1) e^{-\lambda(\alpha_2 - \alpha_1)rt}} = \alpha_2 + \frac{\alpha_1 - \alpha_2}{1 - \left(\frac{s - \alpha_1}{s - \alpha_2}\right) e^{-\lambda(\alpha_2 - \alpha_1)rt}}$$

For  $j = 0$ , we just need to plug in  $s = 0$ :

$$\mathbf{H}_1^+(r, 0, t) = \alpha_2 + \frac{\alpha_2\alpha_1 - \alpha_2^2}{\alpha_2 - \alpha_1 e^{-\lambda(\alpha_2 - \alpha_1)rt}} = \alpha_2 + \frac{A(r)}{B(r)}.$$

Then

$$\left. \frac{d}{dr} \mathbf{H}_1^+(r, 0, t) \right|_{r=1} = -\frac{\mu^2}{\lambda(\mu - \lambda)} + \frac{A'(1)B(1) - A(1)B'(1)}{B^2(1)},$$

where

$$\begin{aligned}
A(1) &= \frac{\mu}{\lambda} \left(1 - \frac{\mu}{\lambda}\right), \\
B(1) &= \frac{\mu}{\lambda} - e^{(\lambda-\mu)t}, \\
A'(1) &= \frac{\mu}{\mu-\lambda} \left(1 + 2\frac{\mu^2}{\lambda^2} - \frac{\mu}{\lambda}\right), \\
B'(1) &= - \left[ \frac{\mu^2}{\lambda(\mu-\lambda)} + \frac{\lambda}{\mu-\lambda} e^{(\lambda-\mu)t} (1 + 2\mu t) \right].
\end{aligned}$$

For  $j > 0$

$$\frac{\partial^j}{\partial s^j} \mathbf{H}_1^+(r, s, t) = (-1)^{j+1} j! \left[ \frac{\overbrace{\alpha_2(\alpha_1 - \alpha_2) \left(1 - e^{\lambda(\alpha_2 - \alpha_1)rt}\right)^j}^{A(r)}}{\underbrace{\left(\alpha_1 e^{\lambda(\alpha_2 - \alpha_1)rt} - \alpha_2\right)^{j+1}}_{B(r)}} + \frac{\overbrace{(\alpha_1 - \alpha_2) \left(1 - e^{\lambda(\alpha_2 - \alpha_1)rt}\right)^{j-1}}^{C(r)}}{\underbrace{\left(\alpha_1 e^{\lambda(\alpha_2 - \alpha_1)rt} - \alpha_2\right)^j}_{D(r)}} \right].$$

Then

$$\frac{\partial}{\partial r} \frac{\partial^j}{\partial s^j} \mathbf{H}_1^+(r, s, t) \Big|_{s=0, r=1} = (-1)^{j+1} j! \left[ \frac{A'(1)B(1) - A(1)B'(1)}{B^2(1)} + \frac{C'(1)D(1) - C(1)D'(1)}{D^2(1)} \right],$$

where

$$\begin{aligned}
A(1) &= \frac{\mu}{\lambda} \left(1 - \frac{\mu}{\lambda}\right) \left(1 - e^{(\lambda-\mu)t}\right)^j, \\
B(1) &= \left(e^{(\lambda-\mu)t} - \frac{\mu}{\lambda}\right)^{j+1}, \\
C(1) &= \left(1 - \frac{\mu}{\lambda}\right) \left(1 - e^{(\lambda-\mu)t}\right)^{j-1}, \\
D(1) &= \left(e^{(\lambda-\mu)t} - \frac{\mu}{\lambda}\right)^j, \\
A'(1) &= \left(1 - e^{(\lambda-\mu)t}\right)^{j-1} \left[ \frac{\mu}{\mu-\lambda} \left(1 + 2\frac{\mu^2}{\lambda^2} - \frac{\mu}{\lambda}\right) \left(1 - e^{(\lambda-\mu)t}\right) + j2\frac{\mu^2}{\lambda} e^{(\lambda-\mu)t} t \right], \\
B'(1) &= (j+1) \left(e^{(\lambda-\mu)t} - \frac{\mu}{\lambda}\right)^j \left[ \frac{\lambda}{\mu-\lambda} e^{(\lambda-\mu)t} (1 + 2\mu t) + \frac{\mu^2}{\lambda(\mu-\lambda)} \right], \\
C'(1) &= \left(1 - e^{(\lambda-\mu)t}\right)^{j-2} \left[ \frac{\lambda^2 + \mu^2}{\lambda(\mu-\lambda)} \left(1 - e^{(\lambda-\mu)t}\right) + (j-1)e^{(\lambda-\mu)t} 2\mu t \right], \\
D'(1) &= j \left(e^{(\lambda-\mu)t} - \frac{\mu}{\lambda}\right)^{j-1} \left[ \frac{\lambda}{\mu-\lambda} e^{(\lambda-\mu)t} (1 + 2\mu t) + \frac{\mu^2}{\lambda(\mu-\lambda)} \right].
\end{aligned}$$